(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
21 July 2005 (21.07.2005)

PCT

(10) International Publication Number
WO 2005/067153 A1

(51) International Patent Classification[7]: H03M 7/30, 7/40, G06F 17/30

(21) International Application Number:
PCT/IB2004/052842

(22) International Filing Date:
17 December 2004 (17.12.2004)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
200310124520.5
30 December 2003 (30.12.2003) CN

(71) Applicant *(for all designated States except US)*: **KONINKLIJKE PHILIPS ELECTRONICS N.V.** [NL/NL]; Groenewoudseweg 1, NL-5621 BA Eindhoven (NL).

(72) Inventor; and

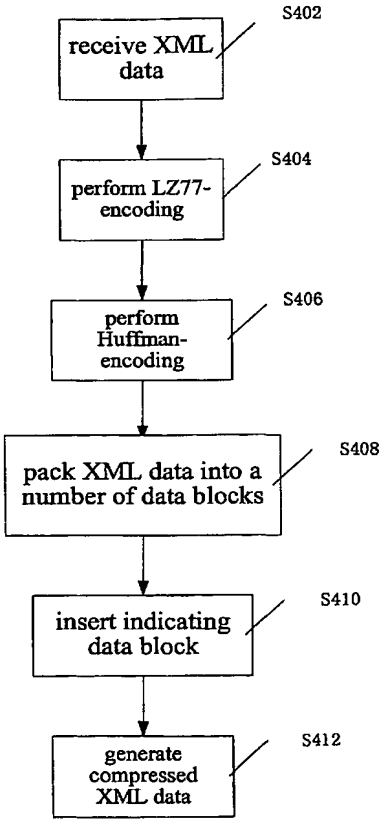
(75) Inventor/Applicant *(for US only)*: **MOREL, Anthoy** [FR/CN]; Philips Electronics China, 21/F Kerry Office Building 218 Tian Mu, Xi Road, Shanghai 200070 (CN).

(74) Common Representative: **KONINKLIJKE PHILIPS ELECTRONICS N.V.**; c/o HAQUE, Azir, Philips Electronics China, 21/F Kerry, Office Building, 218 Tian Mu Xi Lu Road, Shanghai 200070 (CN).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NA, NI, NO, NZ, OM, PG, PH, PL, PT, RO, RU, SC, SD, SE, SG, SK, SL, SY, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, YU, ZA, ZM, ZW.

[Continued on next page]

(54) Title: RAPIDLY QUERYABLE DATA COMPRESSION FORMAT FOR XML FILES



(57) Abstract: A method and device for XML compression with easy querying are provided. An XML file is parsed with a SAX-parser, useless characters such as tabulators and white spaces are removed, indicating data marks are inserted, LZ-77 compression is applied, and finally the data are Huffman-encoded and packed in data blocks. The indicating marks are used to search in the compresed file for tags or literals in the document, based e.g. on alphabetical order. The indicating marks consist of a special character such as a tab and an XML comment; hence they are XML-compatible. The organization of the compressed file in independent data blocks facilitates rapid querying and partial decompression of the compressed file.

WO 2005/067153 A1

(84) **Designated States** *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HU, IE, IS, IT, LT, LU, MC, NL, PL, PT, RO, SE, SI, SK, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Declaration under Rule 4.17:

— *as to applicant's entitlement to apply for and be granted a patent (Rule 4.17(ii)) for the following designations AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NA, NI, NO, NZ, OM, PG, PH, PL, PT, RO, RU, SC, SD, SE, SG, SK, SL, SM, SY, TJ, TM, TN, TR, TT, TZ, UA, UG, UZ, VC, VN, YU, ZA, ZM, ZW, ARIPO patent (BW, GH, GM, KE, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HU, IE, IS, IT, LT, LU, MC, NL, PL, PT, RO, SE, SI, SK, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG)*

Published:

— *with international search report*

— *with amended claims and statement*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.